Project Proposal for IAL620-02: Text Mining & Natural Language Processing

Water Reports & News Classification using Machine Learning models

Pradnya Patil
Master's in informatics and Analytics
University of North Carolina, Greensboro, USA
papatil@uncg.edu

Problem Statement:

Now a days due to global warming and high amount of pollution, we are getting new information about environment and that is accessible to public on every second, the information such as reports, news, books, articles, etc. Using the water report and news, we can classify the data into different category and make the decision more efficiently.

In this project, we seek to answer the following data research questions:

A. Exploratory Data Analysis on the raw downloaded dataset along with text cleaning and transformation.
B. Explain different methods to analyze text and extract features that can be used to build a classification model
C. Classify water news and Reports based on subject text using Machine Learning Models (Decision Trees, SVM, Random Forest...).
D. Predict the probability of the different model of classification.
E. Do sentiment analysis of news and reports related to the water problems based on environmental issue.

Introduction:

In last years, water quality has been threatened by various pollutants and some environmental issues. Therefore, modeling and predicting water reports and news have become very important in controlling water pollution. In this project, models are built to classify the reports and news based on the 5 different target variables using the machine learning algorithms.

For this project, I am collecting the from https://www.kaggle.com/vbmokin/nlp-reports-news-classification. From this analysis, I will learn the key concept of the NLP mostly TEXT classification where the problem of assigning the news and report into their categories. Along with the Classification, I will learn the sentiment analysis also.

From this process, we can tell the reports, or the news are related to environment problems, pollution, treatment, climate indicators or the biotic monitoring in water or in a river basin.

Data Source and description:

The dataset contains 2 files so far - an English-language dataset from the English-language edition of the book, where the co-author, and a Ukrainian-language dataset from a separate Ukrainian-language edition of this book. These datasets contain approximately 95% of the same information.

- Text - One or more sentences from reports or news and 5 binary target features:
- Env_problems - Is the text about an environmental problem? (0 or 1)
- Pollution - Is the text about environmental pollution? (0 or 1)
- Treatment - Is the text about treatment plants or environmental technologies? (0 or 1)
- Climate - Is the text about climatic indicators? (0 or 1)
- Biomonitoring - Is the text about biological, biotic monitoring in water or in a river basin? (0 or 1)

References:

- https://www.kaggle.com/vbmokin/nlp-reports-news-classification
- https://towardsdatascience.com/text-analysis-feature-engineering-with-nlp-502d6ea9225d